SIT743 Multivariate and Categorical Data Analysis
Assignment-1

Total Marks = 120, Weighting - 25%

Due date: 25 August 2019 by 11.30 PM

INSTRUCTIONS:

- For this assignment, you need to submit the following **THREE** files.

  1. **A written document** (*A **single pdf only***) covering all of the items described in the questions. All answers to the questions must be written in this document, i.e, **not** in the other files (code files) that you will be submitting. ***All the relevant results (outputs, figures) obtained by executing your R code must be included in this document.***
     *For questions that involve mathematical formulas, you may write the answers manually (hand written answers), **scan it to pdf** and **combine with your answer document**. Submit a combined **single pdf** of your answer document.*

  2. A **separate** ".R" file or '.txt' file containing your code (R-code script) that you implemented to produce the results. Name the file as "*name-StudentID-Ass1-Code.R*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

  3. **A data file** named "*name-StudentID-TIMyData.txt*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

- All the documents and files should be submitted (uploaded) via *SIT 743 Clouddeakin Assignment Dropbox* by the due date and time.
- **Zip files are NOT accepted**. All three files should be uploaded **separately** to the CloudDeakin.
- E-mail or manual submissions are **NOT** allowed. Photos of the document are **NOT** allowed.
- The questions Q2 and Q3 **do not** require any R programming.

===

Some of the questions in this assignment require you to use the "**ThursdayIsland**" dataset. This dataset is given as a CSV file, named "**ThursdayIsland.csv**". You can download this from the Assignment folder in CloudDeakin. Below is the description of this dataset.

ThursdayIsland dataset:
This dataset gives the air and sea water measurements collected at ***Thursday Island,*** which is an island of the Torres Strait Islands archipelago (North Queensland, Australia)**.**
[http://weather.aims.gov.au/#/station/921].

The data gives 10 minutes sample measurements collected over a 3 month period between March 2018 and June 2018.

The variables include the following (6 variables):

Air Pressure: pressure measurements expressed in units of Hectopascals

Air Temperature: Air temperature in degrees Celsius.

Humidity: Humidity in percentage.

Wind Speed: Wind speed in kilometre per hour

Water Pressure: Water pressure (at 7m below the sea surface) in decibar.

Water Temperature: Water temperature (at 7m below the surface of the sea water) in degrees Celsius.

Q1) [19 Marks]:

- Download the *ThursdayIsland* data file "**ThursdayIsland**.csv" and save it to your R working directory.
- Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.csv("ThursdayIsland.csv", header = TRUE,
sep = ","))
```

- Generate a sample of 5000 data using the following:

```
my.data <- the.data [sample(1:15000,5000),c(1:6)]
```

Save "my.data" to a text file titled "*name-StudentID-TIMyData.txt*" using the following R code (**NOTE: you 'must' upload this data text file and the R code along with your submission. If not, ZERO marks will be given for this whole question**).

```
write.table(my.data,"name-StudentID-TIMyData.txt")
```

Use the sampled data ("my.data") to answer the following questions.

1.1) Draw histograms for 'Humidity' and 'Wind Speed" values, and comment on them. [2 Marks]

1.2) Draw a parallel Box plot using the two variables; 'Air Temperature' and the 'Water Temperature'.
Find five number summaries of these two variables.
Use both five number summaries and the Boxplots to compare and comment on them. [5 Marks]

1.3) Which summary statistics would you choose to summarize the center and spread for the "Air Pressure" data? Why (support your answer with proper plot/s)? Find those summary statistics for the "Air Pressure" data.
[4 Marks]

1.4) Draw a scatterplot of ‘‘Air Temperature’ (as x) and ‘Water Temperature’ (as y) for the ***first 1000 data vectors selected from the “my.data”*** (name the axes).
Fit a linear regression model to the above two variables and plot the (regression) line on the same scatter plot.
Write down the linear regression equation.
Compute the *correlation coefficient* and the *coefficient of Determination*.
Explain what these results reveal. [8 Marks]

Q2) [21 Marks]

2.1) The table shows results of a survey conducted about the type of powerboats (in hundreds) people own, for recreation purposes, in different states over some period in 2019.

		State		
		New south Wales (N)	**Victoria (V)**	**Queensland (Q)**
Powerboat type	**Cruiser (C)**	800	1000	1800
	Jet boat (J)	1200	1200	2000

Suppose we select a person at random,

a) What is the probability that the person owns a jet boat (J)? [1 mark]

b) What is the probability that the person owns a cruiser (C) and from New South Wales (N)? [1 Mark]

c) What is the probability that the person owns a jet boat (J) given that he/she is from Victoria (V)? [2 Marks]

d) What is the probability that the person, who owns a cruiser (C) is from Queensland (Q)? [2 Marks]

e) What is the probability that the person is from Victoria (V) or owns a jet boat (J)? [3 Marks]

f) Find the marginal distribution of the powerboat type. [2 marks]

g) Are powerboat type and state mutually exclusive? Explain [2 Marks]

h) Are powerboat type and state independent? Explain [3 marks]

2.2) There are two coloured boxes kept in a room. The first box is green in colour, and it contains five red and four white balls in it. The second box is black in colour, and it contains three red and four white balls in it. A child is asked to select a box randomly and chose a single ball from the selected box. It is known that this child prefers green colour than black colour, so he prefers to choose green colour 70% of the times. If the ball selected by the child from the box turns out to be white, what is the posterior probability that the white ball came from the black box? **[5 Marks]**

Q3) [5 Marks]

3.1) State two differences between frequentist way and the Bayesian way of estimating a parameter [2 marks]

3.2) Why conjugate priors are useful in Bayesian statistics? [1 mark]

3.3) Give two examples of Conjugate pairs (i.e., give two pairs of distributions that can be used for prior and likelihood) [2 marks]

Q4) Frequentist and Bayesian estimations [31 Marks]

A company called, DataGrind Ltd. (DataGrind in short), has recently created an in-house data science centre. It employed several data scientists to develop various machine learning algorithms to improve their existing processes, and discover new insights from the vast amount of customer and process data that they have collected over several years of operation. During the operation over last year, the company has discovered that the time consumption for debugging the algorithms are contributing significantly in terms of timely delivering a solution to the customers. As part of their planning, the DataGrind wants to model the behaviour, and estimate a parameter θ, which is the probability that a data scientist will find a bug in the algorithm every time he/she compiles the program. Let x_i be the number of attempts (finding a bug, fixing and successfully recompiling a program) a data scientist have tried before successfully finishing his/her algorithm (bug free) over a given period of time (for example, within a work day). Assume that the *number of attempts x_i before success* for a data scientist can be modelled using the following distribution (a ***special form*** of geometric distribution) with an unknown parameter θ as shown in the equation.

$$\boldsymbol{x_i \sim Att(\theta)}$$

$$\boldsymbol{Att(\theta) = p(x_i|\theta) = (1-\theta)^{x_i}\,\theta}$$

Assume that there are N data scientists involved in developing different algorithms (and each person is working on one algorithm over the considered period), and the number of attempts before successfully completing their algorithms by the different data scientists are independently and identically distributed (iid).

4.1) DataGrind first decided to use a frequentist approach to estimate the θ.

a) Show that the joint distribution of number of attempts by N data scientists can be given by the below equation (show the steps clearly).

$$\boldsymbol{p(X|\theta) = \theta^N (1-\theta)^S}, \quad \text{where } \boldsymbol{S = \sum_{i=1}^{N} x_i}$$

[3 marks]

b) Find a simplified expression for the log-likelihood function $\boldsymbol{L(\theta)} = \mathbf{ln}(\boldsymbol{p(X|\theta)})$ **[3** marks**]**

c) Show that the Maximum likelihood Estimate ($\hat{\theta}$) of the parameter θ is:

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{1}}{\mathbf{1} + \overline{\boldsymbol{X}}} \ , \quad where \ \overline{\boldsymbol{X}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{i=1}^{N} \boldsymbol{x_i}$$

[4 Marks]

d) Suppose that over a one day period, the number of attempts that has been observed from 5 data scientists are {2, 0, 4, 1, 3}. What is the Maximum likelihood Estimate $\hat{\theta}$ (MLE) of parameter θ given this data? [2 Marks]

e) Hence, what is the probability that a data scientist will complete the algorithm successfully (bug free) by the end of the work day, assuming that each bug fix takes (a constant) 2 hours, and a work day has 8 hours? [3 Marks]

4.2) DataGrind has now consulted another software development company, called ***SoftExpert***, which has more experience, and obtained some prior information about the probability of finding a bug by a typical programmer every time he/she compiles a program. The ***SoftExpert*** mentioned that their θ value follows a Beta distribution, ***Beta (a,b)***, where a and b are the hyper-parameters of the Beta distribution, with $a = 1$ and $b = 10$.

$$\boldsymbol{Beta(a,b)} = \boldsymbol{K} \ \boldsymbol{\theta}^{\boldsymbol{a-1}} (\mathbf{1} - \boldsymbol{\theta})^{\boldsymbol{b-1}} \quad \textbf{,} \text{ Where } K \text{ is a constant.}$$

a) DataGrind has decided to use this prior information for their estimation. If it uses the Beta distribution prior, **Beta (a,b)**, obtain **an expression** for the **posterior distribution** (show all the steps)**.** Show that the posterior distribution is also a Beta distribution, **Beta (a', b')**, with different hyper-parameters a' and b'. Express a' and b' **in terms of $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{N}$ and $\boldsymbol{S}$**. [5 Marks]

b) Use the values for a and b hyper-parameters suggested by the *SoftExpert****,*** *and* the number of attempts that has been observed from 5 data scientists: {2, 0, 4, 1, 3}, to find the value of a' and b'. What is the posterior mean estimate of θ **?** [4 Marks]

c) Write a R program and plot the obtained likelihood distribution, the prior distribution and the posterior distribution on the same graph. Use different colors to show the distributions on the plot. [3 Marks]

d) Find the posterior mean estimate of θ, if a uniform distribution is used as the prior. Compare the results obtained with the result obtained for Q4.1.d above and comment. [4 Marks]

Q5) Bayesian inference for Gaussians (unknown mean and known variance) [15 marks]

A random sample of n Australian Barramundi fish are caught and measured from the Gladstone region, in Queensland, and their lengths are noted. The average length of the n fish is 80cm. Assume that the lengths of the Barramundi fish are normally distributed with unknown mean θ and known standard deviation 5 cm. Suppose your prior distribution for θ is normal with mean 100 cm and standard deviation of 10 cm.

a) Find the posterior distribution for θ in terms of n. (Do not derive the formulae) [3 Marks]

b) For n=10, find the mean and the standard deviation of the posterior distribution. Comment on the posterior variance [3 Marks]

c) For n=200, find the mean and the standard deviation of the posterior distribution. Compare with the results obtained for n=10 in the above question Q5(b) and comment. [3 Marks]

d) Assume that the **prior** distribution is **changed**, and now the prior is distributed as a triangular distribution defined over the range 65 to 95, as shown below:

$$P(\theta) = \begin{cases} \frac{1}{225}\theta - \frac{13}{45} & for \quad 65 \le \theta \le 80 \\ -\frac{1}{225}\theta + \frac{19}{45} & for \quad 80 < \theta \le 95 \end{cases}$$

Write a R program to implement this triangular prior, and compute the posterior distribution considering $n = 1$. Using R program find the posterior mean estimate of θ. Sketch, on a single coordinate axes, the prior, likelihood and the posterior distributions obtained. [6 Marks]

(**Hint**. Use 'Bolstad' package in R to perform this.
library(Bolstad)
#https://cran.r-project.org/web/packages/Bolstad/Bolstad.pdf)

Q6) Clu**stering: [11 marks]**

6.1) ***K-Means clustering:*** Use the data file "SITEdata2019Aug.txt" provided in CloudDeakin for this question. Load the file "SITEdata2019Aug.txt" using the following:

```
zz<-read.table("SITEdata2019Aug.txt")

zz<-as.matrix(zz)
```

a) Draw a scatter plot of the data. [1 mark].

b) State the number of classes/clusters that can be found in the data (by visual examination of the scatter plot) [1 marks].

c) Use the above number of classes as the k value and perform the k-means clustering on that data. Show the results using a scatterplot (show the different clusters with different colours). Comment on the clusters obtained. [2 Marks]

d) Vary the number of clusters (k value) from 2 to 20 in increments of 1 and perform the k-means clustering for the above data. Record the *total within sum of squares (TOTWSS)* value for each k, and plot a graph of TOTWSS verses k. Explain how you can use this graph to find the correct number of classes/clusters in the data. [3 marks]

6.2) ***Spectral Clustering:*** Use the same dataset (`zz`) and run a spectral clustering (use the number of clusters/centers as 4) on it. Show the results on a scatter plot (with colour coding). Compare these clusters with the clusters obtained using the k-means above and comment on the results. [4 Marks]

Q7) [18 Marks]

For this question you will be using "**ThursdayIsland**" dataset. This dataset is given as a CSV file, named "**ThursdayIsland.csv**". You can download this dataset from the Assignment folder in CloudDeakin.

For this question, we consider only the data from one of the variables, namely **"Water Temperature"** (called as 'WT') from this dataset.

Use the following R code to load the whole data for WT variable

```
the.data <- as.matrix(read.csv("ThursdayIsland.csv", header =
                               TRUE, sep = ","))


#extract the 'Water Temperature'values

WTempdata <- the.data[,6]
```

7.1) Provide a time series plot of the WT data (use the index as the time (x-axis)) using R code. [1 Marks]

7.2) Plot the histogram for WT data. Comment on the shape. How many ***modes*** can be observed in the data? [2 Marks]

7.3) Fit a **single Gaussian** model $\mathcal{N}(\mu, \sigma^2)$ to the distribution of the data, where μ is the **mean** and σ is the **standard deviation** of the Gaussian distribution.

Find the maximum likelihood estimate (MLE) of the parameters, i.e., the **mean** μ and the **standard deviation** (σ).

Plot the obtained (single Gaussian) density distribution along with the histogram on the same graph.
[3 Marks]

7.4) Fit a **mixture of Gaussians** model to the distribution of the data using **the number of Gaussians equal to the number of modes** found in the data (in Q7.2 above) . Write the R code to perform this. Provide the **mixing coefficients**, **mean and standard deviation for each of the Gaussians** found. [4 Marks]

7.5) Plot these Gaussians on top of the histogram plot. Include a plot of the combined density distribution as well (use different colors for the density plots in the same graph). [3 Marks]

7.6) Provide a plot of the **log likelihood values** obtained over the iterations and comment on them. [2 Marks]

7.7) Comment on the distribution models obtained in Q7.3 and Q7.4. Which one is better? [1 Marks]

7.8) What is the main problem that you might come across when performing a maximum likelihood estimation using mixture of Gaussians? How can you resolve that problem in practice? [2 Marks]